2015 IEEE International Conference on Robotics and Automation (ICRA)
Washington State Convention Center
Seattle, Washington, May 26-30, 2015

Leveraging Image-based Prior in Cross-Season Place Recognition

Ando Masatoshi Chokushi Yuuto Tanaka Kanji Yanagihara Kentaro

Abstract— **In this paper, we address the challenging problem of single-view cross-season place recognition. A new approach is proposed for compact discriminative scene descriptor that helps in coping with changes in appearance in the environment. We focus on a simple effective strategy that uses objects whose appearance remain the same across seasons as valid landmarks. Unlike popular bag-of-words (BoW) scene descriptors that rely on a library of vector quantized visual features, our descriptor is based on a library of raw image data (e.g., visual experience shared by colleague robots, publicly available photo collections from Google StreetView), and directly mines it to identify landmarks (i.e., image patches) that effectively explain an input query/database image. The discovered landmarks are then compactly described by their pose and shape (i.e., library image ID, and bounding boxes) and used as a compact discriminative scene descriptor for the input image. We collected a dataset of single-view images across seasons with annotated ground truth, and evaluated the effectiveness of our scene description framework by comparing its performance to that of previous BoW approaches, and by applying an advanced Naive Bayes Nearest neighbor (NBNN) image-to-class distance measure.**

I. Introduction

Cross-season place recognition is one of the most challenging tasks in scene recognition (Fig.1). The appearance of a location can vary depending on geometric conditions (e.g., object configuration, and fresh snow cover) and photometric conditions (e.g., illumination). Such changes in appearance lead to difficulties in scene matching, thereby increasing the need for a highly discriminative, compact scene descriptor.

One of the most popular approaches to place recognition is to translate each image into a bag of vector quantized visual features, termed visual words, and then apply document retrieval techniques that are based on the bag-of-words (BoW) document model [1]. Despite its computational efficiency and robustness, this approach suffers from vector quantization errors, and often fails to handle appearance changes across seasons in practice [2]. Thus far, many of successful methods for cross-season place recognition are variants of the SeqSLAM in [2]–[5], which requires view image sequence measurements as the input. In contrast, we consider a single-view based recognition with important applications where a robot's view only sparsely overlaps with pre-mapped view.

In this paper, we address the challenging problem of single-view cross-season place recognition. A new approach is proposed for a compact discriminative scene descriptor that helps in coping with changes in appearance in the environment. We focus on a simple effective strategy that uses objects whose appearances remain the same across seasons as valid landmarks. Unlike popular BoW scene descriptors that rely on a library of vector quantized visual features, our descriptor is based on a library of raw image data (e.g., visual experience shared by colleague robots, and publicly available photo collections from Google StreetView). The library images need not be associated with spatial information such that the viewpoint and orientation are known, nor be necessarily taken in the target environment; thus they are cheaper than database images and readily available. We directly mine the image library to identify landmarks (i.e., image patches) that effectively explain an input query/database image. The discovered landmarks are considered valid if the image pair is consistent in terms of both geometry and photometry. These landmarks are then compactly described by their pose and shape (i.e., library image ID, and bounding boxes (BBs)) and used as a compact discriminative scene descriptor for the input image. We collected a dataset of single-view images across seasons with annotated ground truth, and evaluated the effectiveness of our scene description framework by comparing its performance to that of previous BoW methods, and by applying an advanced Naive Bayes Nearest neighbor (NBNN) image-to-class distance measure.

A. Related Work

Scene descriptors for visual place recognition (VPC) problems have been studied extensively. Local feature approaches such as BoW scene descriptors have been widely studied from various aspects, including self-similarity of images [6], quantization errors [7], query expansion [8], database

Fig. 1. Single-view cross-season place recognition. The appearance of a place may vary depending on geometric (e.g., viewpoint trajectories and object configuration) and photometric conditions (e.g., illumination). Such changes in appearance lead to difficulties in scene matching, and thereby increasing the requirement for a highly discriminative, compact scene descriptor. In this figure, the panels (top-left, top-right, bottom-left, bottom-right) shows visual images acquired in autumn (AU:2013/10), winter (WI:2013/12), spring (SP:2014/4), and summer (SU:2014/7), respectively.

Our work has been supported in part by JSPS KAKENHI Grant-in-Aid for Young Scientists (B) 23700229, and for Scientific Research (C) 26330297.

M. Ando, Y. Chokushi, K. Tanaka and K. Yanagihara are with Graduate School of Engineering, University of Fukui, Japan. tnkknj@u-fukui.ac.jp

augmentation [9], vocabulary tree [10], and global spatial geometric verification as post-processing [11]. As suggested by our experimental results, previous research on cross-season VPC have shown that the BoW scene model is not sufficiently discriminative and often fails to capture the appearance changes across seasons [2].

Global feature approaches such as the GIST feature descriptor [12] (in which a scene is represented by a single global feature vector) are compact and have high matching speeds. In the robot vision community, global feature approaches have been widely used in the context of cross-season VPC [2], [13], [14]. [2] introduces a robust state-of-the-art VPC framework, called SeqSLAM, for cross-season navigation tasks separated by months or years and opposite seasons. However, the above mentioned frameworks rely on image sequence measurement to cope with appearance changes to improve the discriminative power of global features.

Some other works address the cross-season place recognition problem by using multiple different maps for describing different visual appearance across seasons. [3] proposed a robust approach that can identify typical time-varying appearance of an environment from different databases, with the number of map databases required tending to a constant. [15] presented a framework that uses the stored distinct visual appearances of a workspace, to improve place recognition on future visits, and introduced a novel introspective process, executed between sorties. However, direct memorization of multiple varieties of map databases is required in this framework.

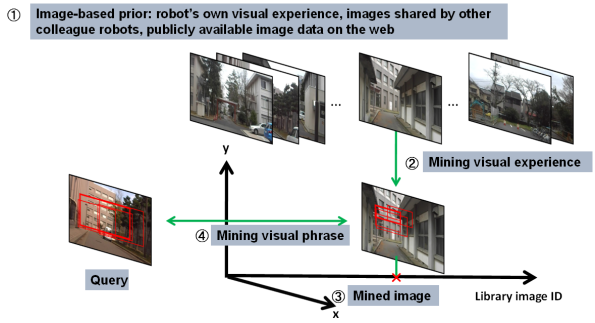
The main contribution of this paper is in presenting a practical recognition approach for single-view cross season place recognition. The landmark verification subsystem employed in Section II-B is inspired by the authors' IROS14 paper [16]. A significantly extended version of the dataset presented in the authors' PPNIV14 paper [17] was used, and this is the first dataset to cover the four seasons.[1]

II. Visual Place Recognition Framework

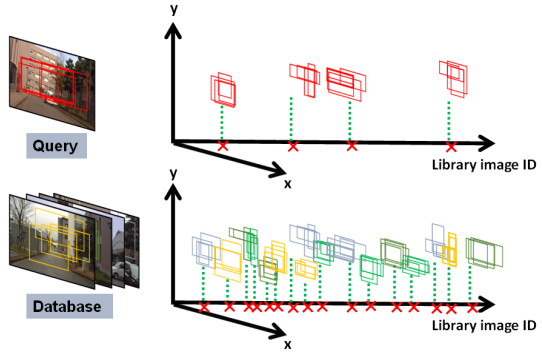
The VPC framework consists of three main steps: (1) landmark proposal, (2) landmark verification, and (3) landmark retrieval (Fig.2). The first step proposes plausible hypotheses of landmark within the input query image. The second step verifies each of the proposed landmarks by mining the library of images to find similar visual patterns. If a landmark proposal that is consistent with a library image in terms of both geometry and photometry exists, the landmark is considered valid and translated into a scene descriptor. Descriptors are also computed for all images in the map database. The third step retrieves the database using the scene descriptors as query. These subtasks are respectively detailed in the following subsections.

For the above interpretation, we assume that a dictionary or a library of random L view images will be given. The library images need not be associated with spatial information such that the viewpoint and orientation are known. Such images are cheaper than the images with spatial information required by the map database, and are more readily available. For example, they can be publicly available resource image data on the web, such as Google StreetView, or a visual experience obtained by the robot itself in a previous navigation, or shared by other colleague robots via information sharing networks. A small subset of J appropriate library images that are most similar to a given input image are selected and used for interpreting the input image. Our experimental results suggest that high recognition performance tends to be associated with the coverage of the database images by these library images.

To translate a given input image to the scene descriptor, we first perform common pattern discovery (CPD) between an input and the library images to mine a set of visual phrases (VPs), i.e., image patches, that effectively explain the input image. Any CPD algorithm can be adopted, but for our purposes, we utilize the fast and stable randomized visual phrase (RVP) algorithm in [18], which can generally handle scale variations among objects without relying on any image segmentation or region detection. We describe the scene description algorithm in Sections II-A, II-B, II-C. Next, we obtain a scene descriptor, which consists of J pairings of



mining visual phrases for landmark discovery



landmark comparison for scene matching

Fig. 2. System overview: proposing, verifying and retrieving landmarks for cross-season place recognition. The proposed framework consists of three distinct steps: (1) landmarks are proposed by patch-level saliency evaluation (red boxes in "Query"), (2) landmarks are verified by mining the image prior to find similar patterns (red boxes in "Mined image"), and (3) landmarks are retrieved by using the bag-of-bounding-boxes scene descriptors (colored boxes in the bottom figure).

[1]Project page with dataset is available at: http://rc.his.u-fukui.ac.jp/projects.html, "Cross-Season Localization"

- a library image ID (an integer), and
- I visual phrases (BBs on the library image).

Because a BB is a considerably lower-dimensional representation than many existing feature descriptors such as 128 dimensional SIFT vectors, the search for similar BBs to a query BB can be conducted quite quickly. We discuss the database search issue in Section II-C.

A. Landmark Proposal

Landmark proposal aims at proposing plausible hypotheses of landmark objects each of which is a crop of the input query/database image. To this end, we utilize bottom up saliency as a cue, which has proven to be effective for unsupervised object discovery in applications such as semi-supervised labeling and anomaly detection. In particular, we adopt the PCA-based saliency measure in [19], as it provides state-of-the-art accuracy for saliency detection while preserving computation efficiency.

We obtain an initial pixel-level saliency using patches collected from the input image. The distinctiveness of a patch is measured by comparing the patch with all other patches. The measurement process begins by extracting a pool of gray-scale patches (size: 9×9 pixels), each of which is represented by an 81 dimensional patch vector. Among the patch vectors, those that belong to homogeneous regions such as the sky are eliminated a priori because we are interested in only those patches that belong to foreground regions. We eliminate them by oversegmenting the image into superpixels using the SLIC algorithm in [20], computing variance of pixel values for each superpixel, and retaining 25% non-homogeneous superpixels with the highest variance. Patches that belong to non-homogeneous superpixels are used for the evaluation of patch's distinctiveness. For each superpixel, the distinctiveness score is averaged over all the pixels belonging to the superpixel, and the averaged score is reassigned to every pixel in the superpixel.

Then, we randomly sample a number of subimages each of which is no larger than 50% of the area of the input image, compute the sum of saliency scores inside each subimage; and finally output $J = 4$ subimages with the highest saliency score as the proposal of landmark objects. Examples of landmark proposals are shown in Fig.6.

B. Landmark Verification

Landmark verification aims at verifying each landmark proposal (i.e., subimage) in terms of the photometric/geometric consistency between the proposal and the image library. This process consists of two distinct steps: (1) search over the image library to find the most similar $J = 4$ images that explain the input image, and (2) discovery of common visual patterns between the input and the mined library images.

In the former process, the pairwise similarity between the input and a library images is evaluated as the number of similar SIFT matches between the image pair. Approximate near neighbor search (ANN) [21] is used to efficiently search for similar SIFTs to an input query SIFT, followed by a RANSAC step to ensure that the normalized L1-distance between the SIFT descriptor pair is smaller than 0.4.

In the latter process, we adopt the RVP [22] to find common visual patterns between the input and the mined library images. Given a set of pairwise SIFT correspondences, the RVP algorithm efficiently searches for common visual patterns between the image pair. The algorithm consists of an iterative scoring process and post-processing. In each iteration, the library image is independently and randomly partitioned into $M \times N$ non-overlapping rectangular patches, and the similarity between the input image and each patch is evaluated in terms of the distance of BoW vectors between the region pair. To translate a visual feature to a visual word, we run the ANN over the library, followed by a verification step to ensure that the normalized L1-distance between the SIFT descriptor pair is smaller than 0.4. We then assign their feature IDs as the visual words, i.e., multiple visual words per feature. In the implementation, we use the histogram intersection as the pairwise distance measure for BoW histograms. Note that after K times of iterations, we have $M \times N \times K = 32 \times 16 \times 200$ patches, and every pixel belongs to exactly K patches. In the post-processing process, the score of each pixel is obtained as the average of the similarity values of the K patches to which the pixel belongs. The final output of the RVP algorithm is the voting map, whose pixel value represents the likelihood of the target landmark proposed.

We execute the abovementioned RVP algorithm and use the resulting voting map to compute the BB whose sum of scores over all the pixels inside the BB are higher than all other potential BBs. Further, the integral image [23] can be used to efficiently compute the sum of the values in the rectangular regions defined by these BBs. The size of a BB should be sufficiently small so that it can be localized well, and should not exceed 10% of the area of the library image.

C. Landmark Retrieval

The scene descriptor consists of J pairings of a library image ID (i.e., an integer) and a set of I visual phrases (BBs on the library image). A BB carries the appearance information of a VP as it indicates the VP region within the library image. Note that our current implementation ensures that each BB is well localized, i.e., smaller than 10% of the image area, and we have already found that there is no need to penalize the size of the BBs. Let $Overlap(\mathscr{B}_{i,j}, \mathscr{B}'_{i',j'})$ denote the area of overlap between a given BB pair $\mathscr{B}_{i,j}$, $\mathscr{B}'_{i',j'}$ when they belong to the same library image or zero otherwise. A large value for the overlap indicates that the VPs cropped by the BBs are similar between the image pair, and vice versa. By aggregating the VP-level similarity, we obtain the image-level similarity:

$$f_{VP}(\mathscr{I}, \mathscr{I}') = \frac{1}{IJ} \sum_{j=1}^{J} \sum_{i=1}^{I} \max_{i',j'} Overlap(\mathscr{B}_{i,j}, \mathscr{B}'_{i',j'}). \quad (1)$$

Since a BB can be compactly represented by a 4D parameter (a considerably lower-dimensional representation than other local feature descriptors such as 128-dim SIFT vectors), the

search for BBs similar to a query BB can be conducted very rapidly.

The proposed scene descriptor tends to produce less meaningful results when there is no common visual pattern (i.e., VP) between the input and the library scenes. We propose the use of the traditional BoW scene descriptor complementary with the proposed scene descriptor, and a modified image-level similarity:

$$f(\mathscr{I},\mathscr{I}') = C_{VP}\cdot f_{VP}(\mathscr{I},\mathscr{I}') + C_{VW}\cdot f_{VW}(\mathscr{I},\mathscr{I}'), \quad (2)$$

where C_{VP} and C_{VW} denote the weighting coefficients and $C_{VP} \gg C_{VW}$. We use the BoW method in [1] and view its output score (i.e., likelihood) value as the similarity f_{VW}.

III. Experimental Results

We evaluate the performance over several datasets that are collected in different seasons and paths. The dataset used in these experiments consists of collections of view images taken around a university campus, using a hand-held camera as a monocular vision sensor.

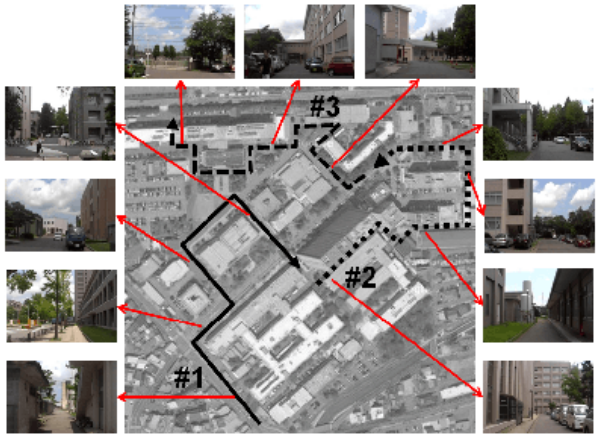


Fig. 3. Experimental environments and viewpoint paths.

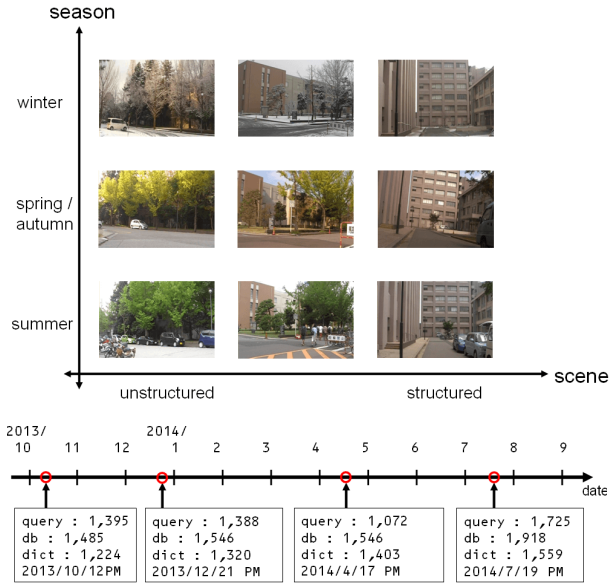


Fig. 4. Datasets. Image datasets are collected for various types of scenes and across seasons (top). The dataset consists of three datasets of query, database, and library images collected during four different seasons over a year (bottom).

A. Settings

Fig.3 and Fig.4 show a bird's eye view of our experimental environments, viewpoint paths, and examples of the dataset. We consider a typical scenario that considers view images that are taken relatively far (1m-5m) from each other [3]. Occlusion is severe in the entire scenes, and people and vehicles are dynamic entities occupying the scene. We traversed each of the three paths #1-#3 three times, collected three independent collections of images and use each for query, library and database image collection. The datasets have been collected across four seasons over a year and cover all the four seasons, as shown in Fig.4. In total, we obtained $3 \times 3 \times 4$ image collections.

B. Map Retrieval Task

Fig.5 shows the input query image, the ground truth image, and the database images top-ranked by the BoW method and by the proposed method. Both the BoW and proposed methods tend to return database images that are similar to the query image to some extent. However, the BoW method tends to fail when there are confusing images in the database whose appearance is partially similar to the query image but with a different structure.

Fig.6 shows example results of proposing, verifying, and retrieving landmarks. For landmark discovery, a set of $J = 4$ library images are selected out of the size $L = 100$ library and $I = 4$ VPs (i.e., landmarks) for each library image are learnt, on the basis of ANN and CPD, in our method, as described in II-B. Discriminative landmarks are successfully found for all the image pairs shown here. However, the reasons for each success vary depending on the content of the input and library images. In the first case, a gate that commonly appears in the input query/database image and a library

Fig. 5. Examples of scene retrievals. From left to right, each panel shows a query image, the ground truth image, the database image top-ranked by the BoW method and by the proposed method.

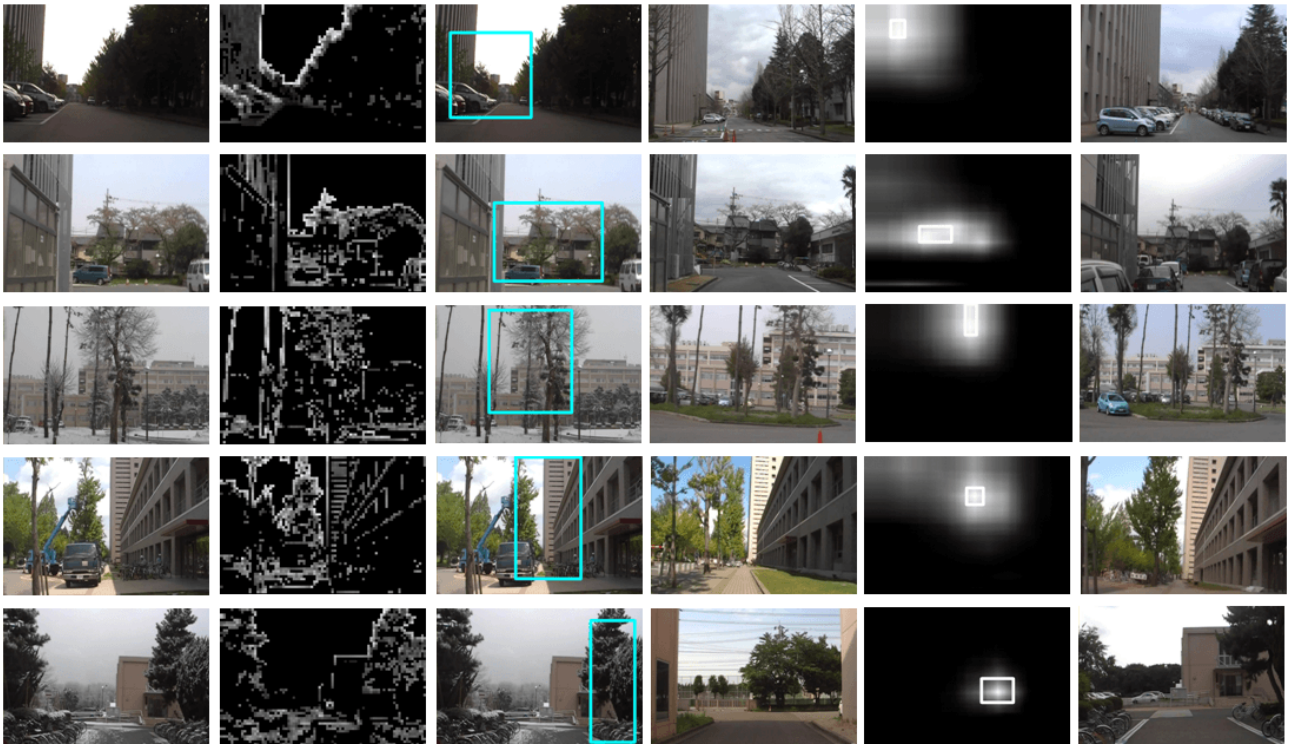


Fig. 6. Examples of proposing, verifying and retrieving landmarks. From left to right, input image, saliency image, landmark proposal (blue bounding box), mined library image, landmark discovered w.r.t. the library image's coordinate, and the top-ranked database image.

TABLE I

SCENE RETRIEVAL PERFORMANCE IN %

Dataset	Query / DB	SP			SU			AU			WI			Avg.
		SU	AU	WI	AU	WI	SP	WI	SP	SU	SP	SU	AU	
FAB-MAP		34.5	34.8	22.1	25.2	31.0	29.8	37.4	30.4	36.4	37.2	41.7	42.0	33.5
VP	#BB=1	27.5	32.6	18.5	25.2	23.9	22.9	31.4	27.1	30.7	29.5	30.8	38.7	28.2
	#BB=2	26.0	31.9	17.6	25.0	20.3	21.2	30.3	26.4	29.8	27.0	29.7	38.4	27.0
	#BB=4	25.4	31.1	**17.1**	24.9	19.8	19.5	29.8	26.4	**28.4**	25.1	31.7	**35.9**	26.3
	#BB=8	**25.0**	**30.9**	17.2	**24.7**	**19.0**	**18.6**	**28.9**	**25.1**	28.5	**24.3**	**28.6**	36.4	**25.6**

image is mined via the landmark discovery. In the second case, there are mainly two dominant objects, a tall building and a small house and between them, the latter is selected as the discriminative landmark, as the appearance of the house has not changed between the input and the library images. In the third case, large portion of the scene is occupied by fresh snow cover, and despite the difficulty, a part of a building is successfully selected as the discriminative landmark. In the fourth case, a part of one of the dominant tall building object is selected as the landmark. In the fifth case, the building object appearing in the library image is not identical to the one appearing in the input scene, but it is selected as a landmark object in our method and successful scene matching is achieved.

C. Performance Results

Table I presents the performance results. We evaluated the proposed visual phrase -based method ("VP") in terms of the retrieval accuracy and compare it with one of the state-of-the-art BoW method ("FAB-MAP") introduced in [1]. For the BoW method, we used the same code as the authors in [1]. A series of independent 200 $\times$ 2 retrievals are conducted for each of the 200 random query images for each two different paths, which respectively consist of path1+path2 and path2+path3 as shown in Fig.3. The retrieval performance was measured in terms of the averaged normalized rank (ANR) in percentage %, which is a ranking-based retrieval performance measure, where a smaller value more favorable. To evaluate ANR, the rank assigned to the ground-truth relevant image was evaluated for each of the 200 independent retrievals, and then the rank was normalized by the database size and averaged over the 200 retrievals. Table I shows that our approach outperformed the BoW method in most of the retrievals considered here.

D. Dependency on Image Prior

Table II presents an investigation of the influence of different settings of image prior on the map retrieval performance. In this study, we are particularly interested in how comprehensively these library images need to cover the path. For example if any building is missing in the library set, it is more difficult for our image prior based method to produce meaningful results. To this end we conducted independent retrieval experiments using different sets of library images, which consist of X[%] samples from the original library and (100-X)[%] samples from another library, for different settings of coverage ratio X = 100%, 75%, 50%, 25%, 0%. The experimental results shown in Table II suggest that a

TABLE II

DEPENDENCY ON IMAGE PRIOR

Dataset	Query / DB	SP			SU			AU			WI			Avg.
		SU	AU	WI	AU	WI	SP	WI	SP	SU	SP	SU	AU	
VP	100%	**25.4**	**31.1**	**17.1**	**24.9**	**19.8**	**19.5**	**29.8**	**26.4**	**28.4**	**25.1**	**31.7**	**35.9**	**26.3**
	75%	29.7	33.4	19.8	25.8	22.9	20.4	31.4	27.6	30.7	27.7	33.7	41.7	28.7
	50%	28.6	32.9	20.6	26.4	24.8	21.7	32.5	28.5	32.3	30.8	34.9	43.5	29.8
	25%	30.7	33.1	21.2	27.5	28.3	25.5	32.8	29.4	31.8	32.9	37.0	43.4	31.1
	0%	32.7	33.9	22.5	27.1	28.9	29.9	35.1	31.6	33.1	38.0	37.5	43.7	32.8

high localization performance tends to be associated with the coverage $X\%$ of the robot's route by these library images. However, the proposed method has a comparable or slightly better performance than BoW even when the coverage is 0%.

E. Frequency of library images

Fig.7 summarizes the frequency of individual library images being used for scene interpretation. In this study, we used all the datasets ({AU,WI,SP,SU} $\times$ paths #1-#3) as test images. As shown in the graph, the frequency is quite different among different datasets, and a small set of library images is more than 10 times frequently used than half of the library images. The most frequently used library images tend to have rich object information while the least frequently used ones tend to have non-distinctive objects such as the sky and trees. The results suggest that a relatively small number of library images would suffice, and intelligent selection of a small number of such useful library images should be addressed in our future work.

F. Comparison Across Seasons

We also investigate the impact of the choice of different season's databases on the retrieval performance. In this study, we use one dataset (e.g., AU) as query images, and for each query image, we perform a pair of independent map retrievals using different season datasets (e.g., WI-SP pair) as database images, evaluate the map retrieval results in terms of the normalized ranks (e.g., r^{WI}, and r^{SP}), and then computes the difference in the map retrieval performance (e.g., $\Delta r = r^{WI} - r^{SP}$). Fig.8 shows the results for different query datasets AU, WI, SP and SU. The performance during the retrieval of SP database using a WI query is higher than during the retrieval of the AU database. The performance during the retrieval of the SU database is higher than during that for the SP database. In addition, performance during the retrieval of the WI database using an SP is higher than that for the AU database. We observe that the retrieval performance tends to be dependent on the appearance similarity between query and library images, as seen in the abovementioned cases.

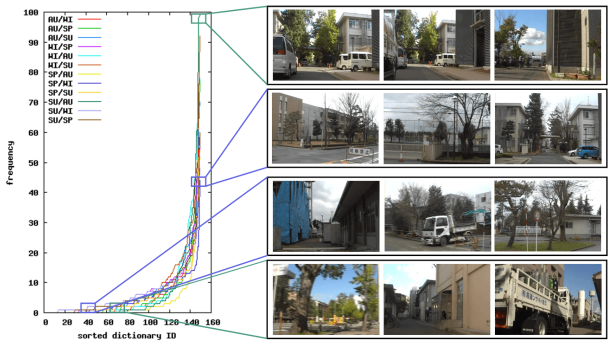


Fig. 7. Frequency of library images.

G. NBNN Image-to-Class Distance Measure

We conducted a final investigation on an alternative image prior based scene comparison scheme, in place of the simple L1 distance and histogram intersection in Section II-B. This study is motivated by the recent success of NBNN image-to-distance measure in [24], which has also been applied to cross domain scene classification tasks [25]. From our viewpoint of localization and mapping, the NBNN measure has several desirable properties: 1) fine vocabulary without relying on vector quantization, 2) lightweight training and efficient classification, and 3) incremental learning of a scene classifier (i.e., perception model) that can be updated by incorporating new training data (i.e., measurements). The NBNN measure works under the following two conditions: 1) raw visual features are used without vector quantization, and 2) image-to-class (rather than image-to-image) distance is used for scene comparison. In the proposed VP framework, condition (1) is satisfied because we do not rely on vector quantization. For condition (2), we view places (i.e., database images) as independent classes and for each class, we prepare a class specific set of training features. The class specific set of training features for each i-th class is obtained by searching through library features for the nearest neighbor (NN) to each feature $j = 1, \cdots, J_i$ in the class (i.e., database image) of interest. In this study, we consider a simple scene

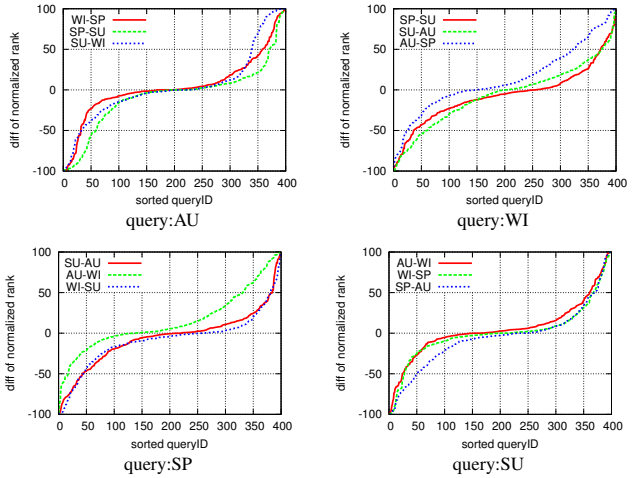


Fig. 8. Comparison across seasons. Horizontal axis: sorted query ID. Vertical axis: difference of normalized ranks $\Delta r = r^{S1} - r^{S2}$ between different season databases. $S1$ and $S2$ are two different seasons indicated in the key (e.g., "WI-SP" indicates $\Delta r = r^{WI} - r^{SP}$).

descriptor that consists of J_i IDs of library features, and not consider the BBs of visual phrases in Section II. In this case, distance between a pairing of query and database scene descriptors is evaluated by

$$dist = \sum_{i=1}^{n} ||q_i - f^{NN}(q_i)||^2, \quad (3)$$

where q_i denotes the features extracted from the query scene image, and $f^{NN}(q_i)$ indicates the corresponding NN library feature that belongs to the database descriptor. The scene descriptor stores only IDs for mapped images to compact the database, and the scene comparison uses both IDs and distance values for query images. Fig.9 shows the results for the NBNN measure. By comparing the data in Fig.9 and Table II, we find that the NBNN measure is effective for many of the cases considered here. However, this comes with a high computational cost. Firstly, by memorizing the scene descriptor that consists of many library features' ID, the same number of SIFT features should be extracted from the database image of interest. Secondly, the NBNN measure requires iterating the NN search for each query feature of interest. One of our future works will be to extend the proposed image prior -based scene descriptor to integrate more robust scene comparison schemes, including NBNN.

IV. Conclusions

The main contribution of this paper is that it addresses the challenging tasks of single-view cross-season place recognition and propose a novel discriminative and compact scene descriptor. In contrast to the widely used BoW scene descriptor that relies on vector quantized feature vectors, our criteria for scene matching is based on raw image matching, which is quantization free. Instead of direct raw image matching between query and database images that is space time intractable, we propose raw image matching between a query/database image and a library of raw image data, such as publicly available image data on the web. We developed a practical place recognition system, by employing efficient and reliable subsystems for raw image matching, including RANSAC geometric verification, common pattern discovery, and approximate near neighbor search. Experimental results show that the proposed framework tends to produce stable recognition results despite the fact that our scene descriptor is significantly space/time efficient.

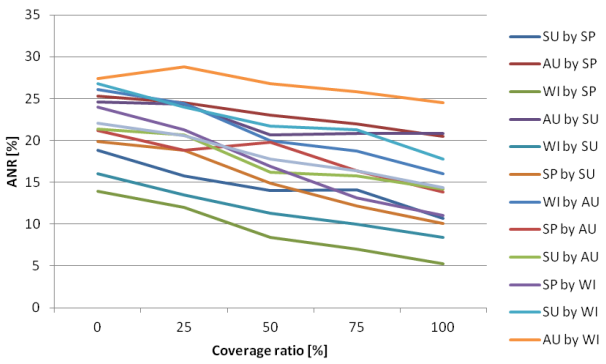


Fig. 9. Results for NBNN image-to-class distance measure.

References

[1] M. Cummins and P. Newman, "Highly scalable appearance-only slam - fab-map 2.0," in *Robotics: Science and Systems*, 2009.

[2] M. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *ICRA*, 2012, pp. 1643–1649.

[3] W. Churchill and P. Newman, "Practice makes perfect? managing and leveraging visual experiences for lifelong navigation," in *ICRA*, 2012, pp. 4525–4532.

[4] M. Milford, "Vision-based place recognition: how low can you go?" *I. J. Robotic Res.*, vol. 32, no. 7, pp. 766–789, 2013.

[5] J. McDonald, M. Kaess, C. D. C. Lerma, J. Neira, and J. J. Leonard, "Real-time 6-dof multi-session visual slam over large-scale environments," *Robot Auton Systems*, vol. 61, no. 10, pp. 1144–1158, 2013.

[6] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *ECCV*, 2010, pp. 748–761.

[7] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012, pp. 2911–2918.

[8] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall ii: Query expansion revisited," in *CVPR*, 2011, pp. 889–896.

[9] P. Turcot and D. G. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 2109–2116.

[10] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *CVPR*, 2007, pp. 1–7.

[11] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007, pp. 1–8.

[12] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *Proc. ACM Int. Conf. Image and Video Retrieval*, 2009, pp. 19:1–19:8.

[13] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," *ICRA Workshop on Long-Term Autonomy held in conjunction*, 2013.

[14] M. Milford and G. Wyeth, "Persistent navigation and mapping using a biologically inspired slam system," *I. J. Robotic Res.*, vol. 29, no. 9, pp. 1131–1153, 2010.

[15] W. Churchill and P. Newman, "Continually improving large scale long term visual navigation of a vehicle in dynamic urban environments," in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, Sept 2012, pp. 1371–1376.

[16] K. Tanaka, Y. Chokushi, and M. Ando, "Mining visual phrases for long-term visual slam," in *IROS*, 2014, pp. 136–142.

[17] M. Ando, Y. Chokushi, and K. Tanaka, "Landmark discovery for single-view cross-season localization," *IROS Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV)*, 2014.

[18] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *CVPR*, 2012, pp. 3100–3107.

[19] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *CVPR*, 2013, pp. 1139–1146.

[20] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *PAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[21] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *Int. Conf. Computer Vision Theory and Application*. INSTICC Press, 2009, pp. 331–340.

[22] Y. Jiang, J. Meng, and J. Yuan, "Randomized visual phrases for object search," in *CVPR*, 2012, pp. 3100–3107.

[23] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, pp. 34–47, 2001.

[24] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *CVPR*, 2008, pp. 1–8.

[25] T. Tommasi and B. Caputo, "Frustratingly easy nbnn domain adaptation," in *ICCV*, 2013, pp. 897–904.